**Abalone Classification with Naive Bayes**

Barışcan Tunalı

Big Data Masters Student

Istanbul Bahcesehir University,

bariscantunali@outlook.com

# ABSTRACT

In this paper, I will explain my study on Abalone classification using Naive Bayes and compare the results between 3 attributes and full attribute train set.

Our problem was predicting categorical value (age group) from continious variables which will be explained later. Age group was grouped by age. Main dataset has no test set so we needed to extract our test set myself.

I applied 2 different set of attributes and 3 different sets of train sets.

# Categories and Subject Descriptors

[**Computing methodologies**]: Machine learning—*Machine learning algorithms; Ensemble methods; Naive Bayes*

**Keywords**

Naive Bayes, Gaussian Distribution, Machine Learning, Data Mining, Classification, Prediction, Abalone

# INTRODUCTION

My first challenge was combining categorical values and continious values in classification.

And another challenge was, actually not a challenge but more of code repeat due to lots of variables to measure everything , implementing Naive Bayes formula into coding and applying 3 different train sets on different validation sets.

The paper is organized as follows: In Section 2, we briefly describe the dataset and data preparation. Next, in Section 3, we explain our classification approach in detail. Experimental results and comparisons with other algorithms are provided in Section 4. Finally, Section 5 concludes the paper.

# DATASET AND FEATURES

Abalone dataset constists of 4177 values. I had 9 attributes, one of them was age.

My attributes are : [Sex], [Length], [Diameter], [Height], [Whole weight], [Shucked weight], [Viscera weight], [Shell weight], [Age]

Attribute name: max value – min value – different values

[Sex] : Categorical : M,F,I

[Length] : 0,815 - 0,075 - 134

[Diameter] : 0,65 - 0,055 - 111

[Height] : 1,13 -0- 51

[Whole weight] : 2,8255 -0,002- 2429

[Shucked weight] : 1,488 -0,001- 1515

[Viscera weight] : 0,76 - 0,0005 - 880

[Shell weight] : 1,005 - 0,0015 - 926

[Age] : 1-3-3

Except sex all of the attributes were continious.

In order to make my program more flexible I created a class and handled all data operations, test train seperations. On presentation layer you only create class and call Initialize() function. After that you only access necessary train and test sets and send them to appropriate classification algorithms.

Abalone100TrainSet = dprp.Abalone100TrainSet;

Abalone100TestSet = dprp. Abalone100TestSet;

NaiveBayes NB1 = new NaiveBayes(); NB1.TrainNB(Abalone100TrainSet);

result=NB1.TestNB(abalone,ParameterType);

ParameterType means using first 3 attributes or using all attributes.

Data preparation class structure as below :

public void Initialize()

{

ExtractData();

TransformData();

SeperateTestData(AbaloneOriginalSetTransformed);

}

*ExcractData()* : Dataset was in .csv format so I made a reader function that reads file, creates class, converts string values to appropriate type and bind them into class. Appends this class object into Abalone List.

*TransformData()* : This function basicly adds only ID into each object. But if we would wanted to apply some data transformation, we would do it here.

*SeperateTestData()* : Gets normalized data and randomly seperates x of data as train and rest as train.

# METHOD

As I mentioned above, I used Naive Bayes to predict continious attributes and used gaussian distribution to predict categorical values. Training Naive bayes and gaussian is bit different from artificial neural networks or similar algorithms. In naive bayes basicly means computing all mean, variance and relevant statistics. These results will be used as Prior to our predicting function.

My Naive Bayes class has 7 functions :

public void TrainNB(List<Abalone> Abalones)

public int TestNB(Abalone Abalone,int ExperimentType)

private void ComputeMeanVariance(List<Abalone> Abalones)

private double ComputeBayesianProbability(double Variance, double ObservationValueOfVariable, double Mean)

private double ComputeEvidenceFor( Abalone abln, bool Sex = false, bool Length = false, bool Diameter = false, bool Height = false, bool Whole\_weight = false, bool Shucked\_weight = false, bool Viscera\_weight = false, bool Shell\_weight = false)

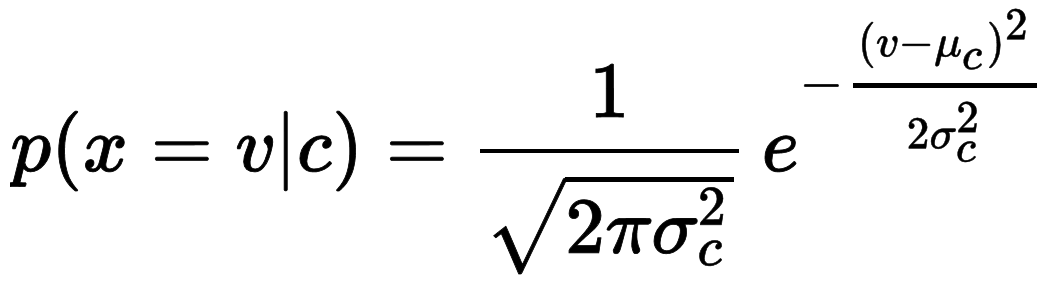
private double ComputeProbabilityFor(int TestAge,Abalone abln, bool Sex = false, bool Length = false, bool Diameter = false, bool Height = false, bool Whole\_weight = false, bool Shucked\_weight = false, bool Viscera\_weight = false, bool Shell\_weight = false)

TrainNB and TestNB are main public functions of the class. TrainNB calls ComputeMeanVariance which calculates all prior statistics.

TestNB has also ExperimentType parameter, which decides we are going to use all attributes or not. Then calls ComputeEvidenceFor and ComputeProbabilityFor functions for young, middle-aged and old. After got the results, it decides which has higher probabity then returns the age category.

ComputeProbabilityFor calls ComputeBayesianProbability to compute NaiveBayes probability. And also computes Gaussian distribution for Sex.

ComputeBayesianProbability implements this function:



GaussianNaiveBayes =

(1 / (2 \* Math.PI\* Variance))\*

Math.Pow(Math.E,

-1\*(Math.Pow(ObservationValueOfVariable - Mean,2) / (2 \* Variance)));

I didn’t use any c# library or dll to implement naive bayes itself but I used lambda expressions and Math library to avoid code complexity.

# EXPERIMENTS

I implemented 9x2=18 experiments, which creates new classes, new train and test sets randomly.

Experiment for 100 train, 4077 Test

Experiment for 100 train, 100 Test(Train set itself)

Experiment for 1000 train, 3177 Test

Experiment for 1000 train, 1000 Test(Train set itself)

Experiment for 2000 train, 2177 Test

Experiment for 2000 train, 2000 Test(Train set itself)

Experiment for 100 train, 4177 Test(All dataset)

Experiment for 1000 train, 4177 Test(All dataset)

Experiment for 2000 train, 4177 Test(All dataset)

And same sets with all attributes

# CONCLUSIONS

Here are 18 experiment results for naive bayes :

3 Parameter Results – Sex, Lenght, Diameter :

EXPERIMENT 1 WITH SEX, LENGHT, DIAMETER VAL SET

Pr \ Re Young Middle Old

Young 84 0 0

Middle 731 2319 943

Old 0 0 0

Correct Predicted : 2403

False Predicted : 1674

Total Predicted : 4077

Young - Real : 815 Predicted : 84

Middle - Real : 2319 Predicted : 3993

Old - Real : 943 Predicted : 0

Total Experimented - Real : 4077 Predicted : 4077

Rate : 0,589404

EXPERIMENT 1 WITH SEX, LENGHT, DIAMETER TEST SET

Pr \ Re Young Middle Old

Young 3 0 0

Middle 21 59 17

Old 0 0 0

Correct Predicted : 62

False Predicted : 38

Total Predicted : 100

Young - Real : 24 Predicted : 3

Middle - Real : 59 Predicted : 97

Old - Real : 17 Predicted : 0

Total Experimented - Real : 100 Predicted : 100

Rate : 0,62

EXPERIMENT 2 WITH SEX, LENGHT, DIAMETER VAL SET

Pr \ Re Young Middle Old

Young 69 0 0

Middle 567 1835 706

Old 0 0 0

Correct Predicted : 1904

False Predicted : 1273

Total Predicted : 3177

Young - Real : 636 Predicted : 69

Middle - Real : 1835 Predicted : 3108

Old - Real : 706 Predicted : 0

Total Experimented - Real : 3177 Predicted : 3177

Rate : 0,5993075

EXPERIMENT 2 WITH SEX, LENGHT, DIAMETER TEST SET

Pr \ Re Young Middle Old

Young 36 0 0

Middle 167 543 254

Old 0 0 0

Correct Predicted : 579

False Predicted : 421

Total Predicted : 1000

Young - Real : 203 Predicted : 36

Middle - Real : 543 Predicted : 964

Old - Real : 254 Predicted : 0

Total Experimented - Real : 1000 Predicted : 1000

Rate : 0,579

EXPERIMENT 3 WITH SEX, LENGHT, DIAMETER VAL SET

Pr \ Re Young Middle Old

Young 48 0 0

Middle 387 1241 501

Old 0 0 0

Correct Predicted : 1289

False Predicted : 888

Total Predicted : 2177

Young - Real : 435 Predicted : 48

Middle - Real : 1241 Predicted : 2129

Old - Real : 501 Predicted : 0

Total Experimented - Real : 2177 Predicted : 2177

Rate : 0,5920992

EXPERIMENT 3 WITH SEX, LENGHT, DIAMETER TEST SET

Pr \ Re Young Middle Old

Young 49 0 0

Middle 355 1137 459

Old 0 0 0

Correct Predicted : 1186

False Predicted : 814

Total Predicted : 2000

Young - Real : 404 Predicted : 49

Middle - Real : 1137 Predicted : 1951

Old - Real : 459 Predicted : 0

Total Experimented - Real : 2000 Predicted : 2000

Rate : 0,593

EXPERIMENT 4 WITH ALL WITH SEX, LENGHT, DIAMETER 100 Train

Pr \ Re Young Middle Old

Young 87 0 0

Middle 752 2378 960

Old 0 0 0

Correct Predicted : 2465

False Predicted : 1712

Total Predicted : 4177

Young - Real : 839 Predicted : 87

Middle - Real : 2378 Predicted : 4090

Old - Real : 960 Predicted : 0

Total Experimented - Real : 4177 Predicted : 4177

Rate : 0,5901365

EXPERIMENT 5 WITH ALL WITH SEX, LENGHT, DIAMETER 1000 Train

Pr \ Re Young Middle Old

Young 105 0 0

Middle 734 2378 960

Old 0 0 0

Correct Predicted : 2483

False Predicted : 1694

Total Predicted : 4177

Young - Real : 839 Predicted : 105

Middle - Real : 2378 Predicted : 4072

Old - Real : 960 Predicted : 0

Total Experimented - Real : 4177 Predicted : 4177

Rate : 0,5944458

EXPERIMENT 6 WITH ALL WITH SEX, LENGHT, DIAMETER 2000 Train

Pr \ Re Young Middle Old

Young 97 0 0

Middle 742 2378 960

Old 0 0 0

Correct Predicted : 2475

False Predicted : 1702

Total Predicted : 4177

Young - Real : 839 Predicted : 97

Middle - Real : 2378 Predicted : 4080

Old - Real : 960 Predicted : 0

Total Experimented - Real : 4177 Predicted : 4177

Rate : 0,5925305

All Parameter Results:

EXPERIMENT 1 WITH SEX, LENGHT, DIAMETER VAL SET

Pr \ Re Young Middle Old

Young 594 215 16

Middle 212 2109 931

Old 0 0 0

Correct Predicted : 2703

False Predicted : 1374

Total Predicted : 4077

Young - Real : 806 Predicted : 825

Middle - Real : 2324 Predicted : 3252

Old - Real : 947 Predicted : 0

Total Experimented - Real : 4077 Predicted : 4077

Rate : 0,6629875

EXPERIMENT 1 WITH SEX, LENGHT, DIAMETER TEST SET

Pr \ Re Young Middle Old

Young 21 4 0

Middle 12 50 13

Old 0 0 0

Correct Predicted : 71

False Predicted : 29

Total Predicted : 100

Young - Real : 33 Predicted : 25

Middle - Real : 54 Predicted : 75

Old - Real : 13 Predicted : 0

Total Experimented - Real : 100 Predicted : 100

Rate : 0,71

EXPERIMENT 2 WITH SEX, LENGHT, DIAMETER VAL SET

Pr \ Re Young Middle Old

Young 351 72 5

Middle 279 1724 745

Old 0 1 0

Correct Predicted : 2075

False Predicted : 1102

Total Predicted : 3177

Young - Real : 630 Predicted : 428

Middle - Real : 1797 Predicted : 2748

Old - Real : 750 Predicted : 1

Total Experimented - Real : 3177 Predicted : 3177

Rate : 0,6531319

EXPERIMENT 2 WITH SEX, LENGHT, DIAMETER TEST SET

Pr \ Re Young Middle Old

Young 118 29 0

Middle 91 552 210

Old 0 0 0

Correct Predicted : 670

False Predicted : 330

Total Predicted : 1000

Young - Real : 209 Predicted : 147

Middle - Real : 581 Predicted : 853

Old - Real : 210 Predicted : 0

Total Experimented - Real : 1000 Predicted : 1000

Rate : 0,67

EXPERIMENT 3 WITH SEX, LENGHT, DIAMETER VAL SET

Pr \ Re Young Middle Old

Young 240 54 4

Middle 202 1206 471

Old 0 0 0

Correct Predicted : 1446

False Predicted : 731

Total Predicted : 2177

Young - Real : 442 Predicted : 298

Middle - Real : 1260 Predicted : 1879

Old - Real : 475 Predicted : 0

Total Experimented - Real : 2177 Predicted : 2177

Rate : 0,6642168

EXPERIMENT 3 WITH SEX, LENGHT, DIAMETER TEST SET

Pr \ Re Young Middle Old

Young 218 44 1

Middle 179 1073 484

Old 0 1 0

Correct Predicted : 1291

False Predicted : 709

Total Predicted : 2000

Young - Real : 397 Predicted : 263

Middle - Real : 1118 Predicted : 1736

Old - Real : 485 Predicted : 1

Total Experimented - Real : 2000 Predicted : 2000

Rate : 0,6455

EXPERIMENT 4 WITH ALL WITH SEX, LENGHT, DIAMETER 100 Train

Pr \ Re Young Middle Old

Young 615 219 16

Middle 224 2159 944

Old 0 0 0

Correct Predicted : 2774

False Predicted : 1403

Total Predicted : 4177

Young - Real : 839 Predicted : 850

Middle - Real : 2378 Predicted : 3327

Old - Real : 960 Predicted : 0

Total Experimented - Real : 4177 Predicted : 4177

Rate : 0,664113

EXPERIMENT 5 WITH ALL WITH SEX, LENGHT, DIAMETER 1000 Train

Pr \ Re Young Middle Old

Young 469 101 5

Middle 370 2276 955

Old 0 1 0

Correct Predicted : 2745

False Predicted : 1432

Total Predicted : 4177

Young - Real : 839 Predicted : 575

Middle - Real : 2378 Predicted : 3601

Old - Real : 960 Predicted : 1

Total Experimented - Real : 4177 Predicted : 4177

Rate : 0,6571702

EXPERIMENT 6 WITH ALL WITH SEX, LENGHT, DIAMETER 2000 Train

Pr \ Re Young Middle Old

Young 458 98 5

Middle 381 2279 955

Old 0 1 0

Correct Predicted : 2737

False Predicted : 1440

Total Predicted : 4177

Young - Real : 839 Predicted : 561

Middle - Real : 2378 Predicted : 3615

Old - Real : 960 Predicted : 1

Total Experimented - Real : 4177 Predicted : 4177

Rate : 0,655255

As we see here normal naive bayes can guess about %55, but full attribute increases it up to %67. Old Abalone prediction rate is very low, almost none. The reason is probably Old Abalone properties are mostly outlier. So in probabilistic calculation even old ones fall into Middle or young.